DOCUMENT RESUME

ED 220 475 TM 820 409

AUTHOR Angoff, William H.; Schrader, William B.
TITLE A Study of Hypotheses Basic to the Use of Rights and Formula Scores. Phase I--Based on Experimental Administration of College Board Tests [and] Phase II--Based on Operational Administration of the GMAT.
PUB DATE 20 Mar 82
NOTE 42p.; Paper presented at the Annual Meeting of the American Educational Research Association (66th, New York, NY, March 19-23, 1982). Some tables in handouts are marginally legible due to small print.

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Comparative Analysis; Equated Scores; *Guessing (Tests); Higher Education; Response Style (Tests); *Scoring Formulas; Secondary Education; *Testing Problems; *Test Items; *Test Theory
IDENTIFIERS Graduate Management Admission Test; Invariance Principle; *Rights and Formula Scoring; Scholastic Aptitude Test

ABSTRACT

In a study to determine whether a shift from Formula scoring to Rights scoring can be made without causing a discontinuity in the test scale, the analysis of special administrations of the Scholastic Aptitude Test and Chemistry Achievement Test and the variable section of an operational form of the Graduate Management Admission Test (GMAT) is discussed. Rights scoring counts correct test items, and Formula scoring involves a penalty for wrong responses. In the two study phases an Invariance Hypothesis was supported that Formula scores are invariant with respect to directions for guessing, indicating its usefulness in equating scores during the period when changed directions are instituted. The effects on examinee behavior of changes in scoring and guessing strategies for scoring consistency are discussed. In the experimental phase the Invariant Hypothesis and a Differential Effects Hypothesis which implies a disadvantage with Formula directions are compared. The effects of the operational phase's realistic conditions to ensure a typical examinee sample and investigation of different item types are discussed. There is an appendix of 18 statistical tables. (CM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED220475

A STUDY OF HYPOTHESES BASIC TO THE USE OF RIGHTS AND FORMULA SCORES

Phase I--Based on Experimental Administration of College Board Tests

William H. Angoff
Educational Testing Service

AERA Meetings
March 20, 1982

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it.

Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Angoff, W.

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 820 109

Introduction

The study that Dr. Schrader and I are presenting this morning was designed and conducted in response to a decision at ETS to prepare for a change in our methods of scoring our answer sheets from Formula to Rights, should the sponsors of the testing programs decide to make that change. In brief, the study was designed to answer the question, Can we develop methods of equating tests that have been administered and scored by Rights to older tests of the same function that had been administered and scored by Formula without causing a discontinuity in the scale?

As background to these presentations it will be useful to state some definitions and to review some of what appear to us to be the more salient issues relevant to our study. First, what we mean by Rights scoring is a simple count of the number of items answered correctly. Correspondingly, what we mean by Formula scoring involves an explicit penalty for wrong responses of the forms, F (Formula) = $R - \frac{W}{k-1}$, where k = no. of response options per item. The intent of the formula is to yield a net score of zero for the aggregate of all items answered at random. It is also understood that in operational practice, directions to examinees will be given appropriate to the method of scoring. That is, that instructions for a Rights-scored test will encourage examinees to respond to every item; instructions for a Formula-scored test will caution examinees that pure guessing is risky and could lower their score. Often added to this instruction is the advice that examinees should eliminate options that they know are clearly incorrect, and to guess from among the others. It is understood, then, that instructions and scoring go hand-in-hand; in operational administrations of a

test the scoring must be performed by the method specified in the directions for administering the test.

Although we might take the foregoing rule for granted, not everyone does. For example, some writers have pointed out that Rights and Formula scores are very highly correlated, from which we are given to infer that it matters little whether the papers are scored one way or the other. But, clearly, this correlation is spurious; in their analyses both scores are based on the same set of responses. Further, there is a moral obligation to score the papers in accordance with the instructions; the data make it clear that examinees adopt their strategies for guessing on the basis of their good-faith assumptions that the papers will in fact be scored as they were led to believe they would be.

The literature on the Rights-Formula controversy is voluminous, but it may be useful to summarize the highlights of the opinions and findings related to this issue. The arguments in favor of Rights-instructions-and-scoring, and opposed to Formula-instructions-and-scoring, include the following: (There are other arguments in addition to these, but they are not quite as relevant to the study as the ones I will mention.)

1. It has been argued that Formula-scored tests introduce irrelevant sources of error variance into the test, stemming from differences among examinees with respect to: 1) their ability to understand directions and take appropriate action; 2) their levels of confidence in assessing their own degree of knowledge; 3) their willingness to take a risk; and 4) their levels of confidence in the testing situation.

2. It has been argued that Rights-scored tests are more reliable than Formula-scored tests. (The opposite has also been argued: that Formula-socred tests are more reliable than Rights-scored tests.)

3. It has been asserted that advice given in connection with Formula-scored tests is bad advice; students are better advised to guess, even in a Formula-scored test, because they usually know more about the content of a test than they think they do. An implication of this position is that students tested with Formula directions are put at a disadvantage relative to those tested with Rights directions. This assertion, which will be referred to as the Differential Effects Hypothesis in this paper, is central to the purpose of this study.

4. The Differential Effects Hypothesis also holds that students who are less inclined to take a risk and guess are further disadvantaged by the Formula-directed test.

5. Argument and evidence have been advanced that Rights scores provide discriminations below mean chance, a point which is defined as a zero raw score, in Formula-score terms.

On the other hand, there are arguments in support of Formula and in opposition to Rights:

1. There have been objections to Rights-directed tests on the ground that they encourage indiscriminate random guessing, especially near the end of the test when there is little time to consider the remaining items carefully. Because of the random guessing component, which adds error variance of its own, Rights scores tend to be less reliable and less valid than Formula-scored tests.

2. Although it may be true that some examinees may be better advised to guess in a Formula-scored test than to omit--because they know

more than they think they do--there are other examinees, especially lower-scoring examinees, who do worse, because their information is often misinformation. That this is true is shown by the preponderant numbers of below-chance asymptotes often observed in item response curves.

3. Even when instructed to guess when they are in doubt, some examinees will fail to respond to every item.

4. It has been suggested that Formula scoring may compensate for differences in instructions. Ledyard Tucker once suggested that even when two groups of students (of equal ability) have been given different instructions regarding guessing, Formula-scoring will tend to equalize the scores. This is a major consideration because it implies that, on the average, the difference in Rights scores for the two groups is a difference due to chance responding, that, on the average, students do not know more than they think they do. This assertion, which will be referred to as the Invariance Hypothesis because its claim is that Formula scores are invariant with respect to directions for guessing, is competitive with the Differential Effects Hypothesis. It is the principal purpose of this study to determine which of these hypotheses is more clearly supported by the data.

As I have already indicated, what triggered off this study was the question whether we could shift from Formula to Rights without causing a discontinuity in the scale. The question is more complicated than it may appear, because--also as already indicated--what is involved is not merely a change in scoring but a change in guessing strategies with which the scoring must, in fairness to the examinees, be made consistent.

A concomitant purpose of the study, one of major importance, was to determine whether the support for the Invariance Hypothesis could safely be used as a basis for equating Rights scores to Formula scores. It was also intended that additional studies would be carried out to help clarify some of the issues in the current debate over Rights and Formula scoring--studies of reliability and parallelism, studies of the ability of students to shift from Rights strategies to Formula strategies, and vice versa, and studies of nonresponse behavior by ethnic background and score level.

Study Design

The total study was designed in two phases. The first phase was based on special administrations of the College Board SAT and the College Board Chemistry Achievement Test. The second phase was based on data from experimental sections of the Graduate Management Admission Test (GMAT).

In the first phase of the study special administrations of two forms of the College Board SAT and one form of the College Board Chemistry Achievement Test were held in selected high schools in which random samples of examinees were specially assigned to take the test under different instructions for guessing. About 8000 students participated in this phase, 6300 taking the SAT and 2300 taking the Chemistry Test. In the second phase of the study, quite separate from the first, random samples of examinees who were taking the Graduate Management Admission Test (GMAT) for admission to schools of business and management, were given additional, nonoperational tests under different instructions for guessing.

For convenience in presentation, I will undertake to describe the design, analysis, and results of the SAT/Chemistry Test phase of the study.

Dr. Schrader will describe the design, analysis, and results of the GMAT phase of the study, and will also present our interpretation of the findings and our general conclusions.

The studies of the SAT and the Chemistry Test were each so designed that random subgroups took the same test under different directions. Two forms of the SAT-verbal were used in the analysis, one (Form A) on which most of the analysis was based, and the other (Form B), which was used for confirmatory analyses. Since the SAT is operationally administered in two separately-timed half-hour parts--45 items in the first part and 40 items in the second--it was possible to administer the test to different groups in different patterns of directions as shown in Table 1: Rights directions on Part 1 and Rights on Part 2 to one group; Rights on Part 1 and Formula on Part 2 to a second group; Formula on Part 1 and Rights on Part 2 to a third group; and Formula on Part 1 and Formula on Part 2 to a fourth group. This was the arrangement for Form A of the SAT-verbal. Form B was administered to two additional groups, one taking it with Rights directions for both Parts 1 and 2, and one taking it with Formula directions for both Parts 1 and 2. Six sets of testbooks were prepared, corresponding to the six groups of students just described (and shown in Table 1), with the directions for guessing (Rights or Formula) printed in the testbooks. Because the timing for all six sets was identical, it was possible to administer all six sets to different students in the same testing room at the same time. It also permitted "spiralling" of the testbooks, by which I mean that we were able to package the testbooks in the order 1, 2, 3, 4, 5, 6, 1, 2, 3, 4, ... and distribute them to the examinees in that order. The result of the "spiralling" was that every sixth examinee received the same testbook

and that the six samples of students were very nearly equivalent samples. In fact, under the conditions of the experiment, the samples were more nearly equivalent than would have been achieved by simple random sampling procedures. In addition, this method of achieving equivalent groups was much more practical in an actual test administration of this sort than the use of random sampling procedures.

As a further check on some of the results of the analysis of the SAT, a form of the College Board Chemistry test was administered to a separate sample of students studying chemistry. Here too, different directions for guessing were prepared, one to encourage students to guess (Rights directions) and one to discourage them from guessing (Formula directions). The directions used are shown on the inside cover of the handout. Again, two types of testbooks were printed, each with its own directions, spiralled, and administered to students in each classroom, with the result that alternate students (Group 7) took the test under Rights directions and alternate students (Group 8) took the test under Formula directions. Table 1 of the handout also describes this arrangement.

Sample Characteristics

The sample of students chosen to take the SAT was drawn systematically from a list of high schools with at least 50 College Board candidates each, and with some care to secure a larger than usual participation of Black and other minority students. As a result of this effort, the final SAT sample comprised 19+% Blacks and about 26% minority students generally. The Chemistry sample contained 1.6% Black and 5.9% minority students generally. There were 58% female and 42% male students in the SAT sample, and 47% female and 53% male students in the Chemistry sample.

Results

Effect of Directions on Mean Formula Scores

The first and most important findings are shown in Table 2, which gives mean Rights and mean Formula scores for pairs of groups of examinees who took the tests under the two sets of directions. As expected, the differences in mean Rights scores in each pair are substantially different, ranging from 8 hundredths to 22 hundredths of a standard deviation. However, when the same answer sheets are rescored by Formula, the differences in mean invariably diminish, ranging from zero (or essentially zero) to 5 hundredths of a standard deviation. In effect, students who guess, as they are encouraged to do on Rights-administered tests, do little better, on the average, than students who are discouraged from guessing, as is the case on Formula-administered tests, when the tests are scored by Formula. To say this slightly differently: Contrary to the assertions found in the literature, we find that students gain very little by guessing on a Formula-scored test. These results also cast some doubt on the assertion that students know more then they think they do, or that their guessing is likely to be informed guessing. While such an assertion may conceivably be true for some examinees and for some items, it appears that there are enough examinees and items for which it is not true to nullify the effect over all.

Table 3 examines the difference in Formula scores on Part 2 of the SAT for students stratified by ability on Part 1 and tests these differences for significance in an analysis in which the covariates were: School, Sex, Rank in Class, and Ethnic Group. In another part of the analysis students were stratified by Items Not Attempted. In none of these tests was the difference in adjusted Formula-score means greater than one raw score point, and in only

three instances--when the numbers of cases were 1500 or more--were the differences significant. From these results it is seen that Formula scoring appears to nullify any differences resulting from differential strategies for guessing; the higher mean scores earned by the students taking the Rights-administered test appear to be higher largely because of random guessing. What is of particular significance here is that this conclusion holds at all levels of ability, and for students with different numbers of items left unattempted. Table 4 shows that these results are not characteristic of White students only; similar nonsignificant differences are found for Black students.

Dr. Lyle V. Jones, who had seen these data a few months ago, reworked them to determine whether standardized Black-White differences in mean score, by level, were greater when the tests were administered and scored by Rights than when they were administered and scored by Formula. Although he did find detailed differences between the Rights and Formula conditions, he found no evidence that one of the modes of administration and scoring favors one group over the other. The standardized mean differences for the two ethnic groups are about the same regardless of level of ability and method of administration and scoring.

Table 5 gives the results of an analysis designed to determine whether students who are called on to change their guessing strategies from Part 1 of the test to Part 2 do any worse on Part 2 than students who do not have to change their guessing strategies. The results show clearly that students, both Black and White, effect this change easily; their mean scores on Part 2

are no lower--indeed, slightly higher--than the scores of students whose guessing strategies on Part 1 of the test were the same as those on Part 2.

Effect of Directions on Nonresponse

Tables 6 and 7 give the numbers of items Omitted, Not Reached, Not Attempted (i.e., Omitted plus Not Reached), and Guessed on Part 2, separately by score interval on Part 1 of the test, and within interval, by ethnic group. As expected, the numbers of items Omitted and Not Reached (and Not Attempted) decline, on the average, for all three ethnic groups shown--White, Black, and Black plus Hispanic--with increasing score on an independent, but correlated, variable. Similarly, scores on an index of guessing--arbitrarily developed from a count of the Wrongs diminished by a count of Omits--also decline as a function of increasing score on an independent, but correlated variable. There are, however, no systematic differences in these mean counts across ethnic groups. Table 6 is based on the data for Group 1. Table 7 gives the same kind of information, but for Group 2, and shows the same trends. Corresponding tabulations for Groups 3-6 show similar trends. Note in Table 6 that even when students are encouraged to answer every item they still have some items unmarked--about 2, on the average, in the 40-item test.

Table 8 simply displays the fact that, as measured by the W-O index, students do what they are instructed to do. Guessing is more frequent, as expected, on Rights-directed tests than on Formula-directed tests.

Intercorrelations, Reliabilities, and Parallelism

One would normally expect that, other things equal, correlations between scores on two separately-timed parts of a test will be higher, if the directions for guessing for those two parts are the same, than if the directions for guessing are different. Table 9 supports this hypothesis. As we look down each column of the table we find that the correlation

between Rights scores, between Wrongs scores, and between number of items Omitted are higher when the directions on Parts 1 and 2 are the same than when the directions on the two parts are different. The correlations between the Wrongs on the two parts confirm this observation and show these differences quite clearly. The differences in the correlations of Omits are clearest of all. There we see correlations of .784 and .770 for tests administered with the same directions, as compared with correlations of .246 and .318 for tests administered with different directions.

Confirmation of this observation is seen in Table 10, where in every column the highest correlations between scores on the two parts are those when the directions are the same (rows 1 and 4) than when the directions are different (rows 2 and 3). The figures in Table 10 also indicate that, on average, the correlations between the parts are higher when the scoring methods are the same (columns 1 and 4) than when the scoring methods are different (columns 2 and 3), a finding that is confirmed by the data for Form B. Indeed, the highest correlations in the Form A part of the table are those obtained when the two parts of the test were administered and scored in the same way ($r_{R_1R_2}$ = .822 and $r_{F_1F_2}$ = .824). However, these differences, while remarkably consistent, are not great, confirming the conclusions drawn in previous research--however improperly they were drawn--that the rank ordering of examinees on two tests administered and/or scored by Rights is not much different from the rank ordering when the tests are administered and scored by Formula.

It is noted, by the way, that the data of Table 10 contain parallel-forms reliabilities and permit the comparison of such reliabilities under the two conditions of administration and scoring. The differences between

the reliabilities for these two conditions are very small indeed, but there is a hint in these data that Formula administration and scoring is slightly more reliable than Rights administration and scoring.

Examination of Table 11 tells us that in general, the reliabilities for Part 1 are higher than those for Part 2--a not unexpected finding inasmuch as Part 1 is slightly longer than Part 2. But what is of greater interest is that the internal consistency reliability of Rights scores is greater than the internal consistency reliabilities of Formula scores. This finding is contrary to the parallel-form reliability data observed in Table 10, where the Formula score reliabilities were higher than the Rights score reliabilities. However, again, these differences are very small. On this basis it is reasonable to conclude that considerations of test reliability should not be decisive in choosing one method of administration and scoring over the other.

Table 12 presents evidence on the parallelism of the two conditions of administration and scoring, and tells us that the true-score correlations between the two parts of the test which were administered and scored in different ways--the italicized values of .970 and .975 in columns 2 and 3--are sufficiently similar to the true-score correlations between the two parts when administered and scored in the same way--the italicized values in columns 1 and 4 of .981 and 1.013, essentially unity--to dispel any concerns that they may be measuring substantially different abilities.

Equating

The original purpose of this research, it is recalled, was to test the validity of the Invariance Hypothesis, which is that Formula scoring

renders equivalent the scores of examinees, some of whom have taken a test under Rights directions, while others have taken the test under Formula directions. Although the results of this test of the hypothesis are relevant to a variety of other psychometric issues, they are also important in providing a means for equating tests that have been administered under Rights directions to other tests of the same ability that had been administered under Formula directions.

The particular method of equating that was examined for possible bias resulting from the possible failure of the Invariance Hypothesis was one we refer to as the Invariant Link Method, where one group (α) takes Form x and a link test, Form v, under Rights directions and the other group (β) takes Form y and the same link test, Form v, under Formula directions. Ordinarily, when there are no complications due to differences in test directions, the data for the link test are used to adjust for any differences in mean and variance between the two groups. Under these conditions, however, because Form v is administered to the two groups under different directions for guessing, we know that the examinees' responses to the test questions do not have the same meaning. This is apparent when we examine the figures in Table 8, and the data of Table 2, which make it clear that the means of Rights scores for random groups who took the tests under the two directions are, as expected, quite different. However, we also see in Table 2 that when the papers are rescored by Formula, the means are much more nearly equivalent, as suggested by the Invariance Hypothesis. Therefore, the method of interest is one where the design for administration is as was just described, but where the method of treating the data calls for scoring the link test answer sheets for both groups by Formula. The

conversion equations produced by applying this procedure were compared with the conversion equations resulting from the use of three different criterion equating methods and found to yield results reasonably consistent with what we would have found had these complications in test administration not existed.

Conclusions

The outcomes of these studies of SAT and College Board Chemistry support our view that the Invariance Hypothesis and its implications for the effect on examinee behavior are essentially tenable. They also support our view that the Invariance Hypothesis may be usefully applied in equating Rights scores to Formula scores in testing programs where the decision is made to change from one mode of administering and scoring the tests to the other.

A Study of Hypotheses Basic to the Use of Rights and Formula Scores

Phase II - Based on Operational Administration of the GMAT

William B. Schrader

Educational Testing Service

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Schrader W.

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

AERA Meeting

March 20, 1982

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Study Design

In this experiment the variable section of an operational form of the Graduate Management Admission Test was used to study the effects of Rights and Formula directions. Utilizing these data, taken from the regular administration of the GMAT, offered several advantages over special administrations. First, conducting the study as part of the GMAT Program provided realistic conditions of motivation and ensured that the sample was reasonably typical of GMAT examinees. In addition, the very large sample size made it possible to study five different item types under the two conditions of administration and to have large samples for each of the ten combinations. With this arrangement, it was possible to use data for the operational part corresponding to each item type in the equating analyses. On the other hand, conducting the study as part of the program imposed certain restrictions on the study design. In particular, it was considered essential that the experiment be conducted in such a way that it would have no effect on <u>any</u> examinee's score of record, but with the additional condition that no examinee could infer that the test material was experimental and would not count toward his or her score. I should add that the decision to conduct this study as part of a regular administration was based on a judgment that it would not be feasible to conduct special test administrations for the GMAT study because of the difficulty in identifying and obtaining the

cooperation of suitable experimental samples.

The experimental tests were all given as the last of the eight separately-timed sections at the October 1980 administration of GMAT. Each of five item types was administered as a 30-minute separately-timed part, and each of the resulting five experimental tests was administered with Rights directions to one group of students and with Formula directions to a different group of students. Thus, there were ten different experimental tests. These ten tests were distributed by spiralling to the entire examinee group, with the result that very nearly equivalent groups of examinees took each test. Every examinee took the experimental test after the operational test. The method of administration thus permitted the comparison of equivalent groups of examinees who took the same test under one or the other of the two directions. The format of the operational and experimental tests is shown in Table 13.

Because all examinees who took the experimental tests took the same form of GMAT, and because each of the experimental tests corresponded to one of the parts of the operational test, it was possible to use the operational test part scores in analyzing the experimental test results.

Major concerns in planning the test administration were to provide appropriate Rights- and Formula-scoring directions for the experimental tests, in order to ensure that examinees read the directions for those tests and to ensure that they were taken under normal test-taking conditions.

The Supervisor's Manual for the test asked the supervisor to read aloud the following statement just preceding the administration of the experimental tests:

> During the next 30 minutes you are to work only on Section 8. Read the directions at the beginning of Section 8 carefully and answer the questions in accordance with these directions. Turn to Section 8, read the directions and begin to work.

Because the directions were relatively brief, it was decided that examinees could read the directions for their test within the time limits for the test.

Although the Formula directions were somewhat longer, it was judged that this difference would be offset by the fact that the directions for these experimental tests were the same directions that had been given to the examinees in previous sections. The guessing directions are shown on the page facing Table 13.

Sample Characteristics

The study sample was defined as all examinees for whom scores were reported on the October test and who attempted at least one item on the experimental test. The total sample size was 55,780. The ten sub-samples obtained by spiralling ranged in size from 5,408 for Group 5 and 5,409 for Group 10 to 5,739 for Group 1 and 5,738 for Group 6. Means of GMAT total scores for the ten groups were quite similar, ranging from 473 (Group 6) to 477 (Groups 7 and 8).

The overall mean for the total sample was 475, with a standard deviation of 106. By comparison, GMAT candidates tested from November 1975 through July 1978 had a mean score of 461 with a standard deviation of 107.

Results

Effect of Directions of Mean Formula Scores

On logical grounds, the possible differential effect of directions on Formula scores depends on whether examinees would do better than would be expected by chance on items that they would answer under Rights directions but would not answer under Formula directions. Although it is anticipated that the means of Rights scores would be clearly higher for those tested under Rights directions than for those tested under Formula directions, it would be expected, under the Invariance Hypothesis, that this difference is caused by random responses to items that normally would not be attempted under Formula directions. Under the Invariance Hypothesis, the difference would be greatly reduced, if not caused to vanish entirely, if a correction for guessing, as is normally applied in Formula scoring, is used. Table 14 provides an opportunity to examine the validity of this hypothesis.

As expected, Table 14 shows that, except for the Practical Business Judgment part, the means of the Rights scores are higher for the groups that received Rights directions than for the groups

that received Formula directions. This is not true, however, for the means of the Formula scores. Formula-score means are remarkably similar for the groups receiving the two types of directions on the tests of Reading Comprehension, Problem Solving, and Usage, and reasonably similar on the other two tests. As it happened, the mean Formula score is higher for those tested with Formula directions than for those tested with Rights directions in four of the five comparisons, including of course, the test of Practical Business Judgment. This result is contrary to the hypothesis that examinees would earn better-than-chance scores on items that they would not answer with Formula directions. On the whole, the results support the hypothesis that Formula scores are invariant with respect to directions.

Effect of Directions on Nonresponse

Table 15 shows the mean and standard deviations of items omitted, items not reached, and items not attempted under Rights and Formula directions for each of the five experimental tests. As in Dr. Angoff's paper, items not attempted are separated into two parts -- unanswered items that precede the last item attempted (called "omits") and unanswered items that follow the last item attempted (called "not reached").

For all tests except Practical Business Judgment, all three of the indicators show a higher mean, as expected, under Formula

directions than under Rights directions. The exceptional results for Practical Business Judgment are presumably the result of the fact that there were very few non-attempts even under Formula directions.

For most tests, the difference in mean omits, not reached, and nonattempts is modest. This result probably arises from the fact that the Rights directions, although explicit, were not strongly emphasized or repeated, and from the fact that the Formula directions encourage examinees to answer any items about which they have any knowledge. Of course, failure to answer every question tends to penalize an examinee tested under Rights directions. It would be expected that there would be fewer nonattempts on operational tests administered under Rights directions than we found in this experiment. On the whole, we believe that this limitation would not affect the comparison of mean formula scores under the two directions.

As shown in Tables 16 and 17, there is a clear relationship between scores on the operational test and the means and standard deviations of the four indicators. As score level increases, the means and standard deviations decrease. These trends are reasonable, because high-scoring students would encounter fewer items than low-scoring students on which they would be called upon to guess or to omit. Comparison of Table 16 with Table 17 shows that with very few exceptions, examinees at a particular score level show higher means for each of the four indicators under Formula directions than under Rights directions. This result indicates that directions <u>did</u> affect guessing

strategies. Of course, if students receiving Rights directions had complied fully with the instruction to answer every question, the means for all three indicators of nonresponse would be zero.

In general, the results for the other four part-scores were similar to those for Reading Comprehension, except that for Practical Business Judgment, results for Rights and Formula directions were virtually undistinguishable.

Table 18 shows the results for the guessing index defined as Wrongs minus Omits discussed by Dr. Angoff in his paper. If, as he pointed out, Wrongs may be regarded as an indicator of guessing and Omits are regarded as evidence of refraining from guessing, the difference may be used as an indicator of the effect of directions on the tendency to guess rather than to omit. The results show a modest effect of Formula directions in reducing the tendency to guess for each of the five tests.

Reliability and Parellelism

Can scores earned on a test administered with Formula directions and scoring be regarded as parallel to scores earned on the same test administered with Rights directions and scoring? This question has considerable theoretical interest and also has practical implications for score equating.

Table 19 gives the basic statistical results on this question for the GMAT data. Reliabilities were computed by Kuder-Richardson

Formula No. 20 for Rights scores and by Dressel's adaptation of K-R (20) for Formula scores. Before discussing the results for reliability, I would like to point out that in the column headed "Exp. Scored Formula," the reliabilities for Practical Business Judgment are repeated, to facilitate comparison with results for the two sections of the operational test, for which separate reliability coefficients were determined. Differences between the two coefficients for each operational test are entirely the result of sampling fluctuations because only Formula directions and scoring results are shown for the operational tests. For the experimental tests, main interest centers on the underlined coefficients for each test. Reliabilities tend to be somewhat higher when Rights directions and scoring are used than when Formula directions and scoring are used. However, in this study, the expectation that examinees would answer every question under Rights directions was not fulfilled. To the extent that the unanswered items affected the reliability of Rights scores obtained when Rights directions were used, the comparisons of reliabilities are not rigorous.

Formula scores on the operational tests tend to yield higher correlations with scores on the corresponding experimental test when the experimental test is given with Formula directions and scoring than when the experimental test is given with Rights directions and scoring. For the two quantitative tests, Problem Solving and Data Sufficiency, the difference in correlations is appreciable. The same

pattern is apparent when true scores rather than observed scores are considered. However, the differences between true score correlations for experimental tests given with different directions are generally small when Formula scoring is used for both Formula- and Rights-directed tests.

On the whole, it appears that, under the conditions of this experiment, the effects of different directions and scoring on parallelism as measured by relative true-score correlations are fairly small. Thus it should be possible to equate scores earned before a change in directions with scores earned after the change, although the results for the quantitative tests may be ambiguous on this point.

As in the College Board studies, a number of equatings were performed to find out whether use of the Invariance Hypothesis in equating would yield satisfactory results. It was concluded that these results agreed well enough with equatings using customary methods to justify the use of the Invariance Hypothesis for equating if difference directions were adopted by a testing program.

Implications of College Board and GMAT Studies

The Invariance Hypothesis implies that Rights scores of examinees would be higher under Rights directions than under Formula directions but that with Formula directions, the differences between the groups would be considerably diminished. On the whole, both

the College Board studies and the GMAT studies yielded results that appear to be consistent with this hypothesis. Analyses of SAT-verbal indicated that the Invariance Hypothesis was supported not only for the total group but for groups stratified on the basis of ability. It follows that under the conditions of these studies, the Differential Effects Hypothesis--that students would earn higher Formula scores under Rights directions than under Formula directions--is not supported, on the whole, by the results. These results suggest also that Formula scoring has the effect of compensating for differences in guessing strategies, so that it is not necessary to require every examinee to answer every item in order to equalize guessing strategies. Finally, the results indicate that the Invariance Hypothesis should be useful in equating scores during the period when changed guessing directions were instituted in a program.

Results of the analyses of unattempted items indicated that both guessing directions and ability were related to the number of items that were not attempted, but that minority group membership was not related to guessing behavior when ability was controlled. Under the conditions of the study, examinees did not answer every question under Rights directions, although they were informed that it would be to their advantage to do so.

With respect to score reliability, the parallel-forms reliabilities for SAT-verbal for Rights directions-and-scoring and for Formula directions-and-scoring are virtually equal. The K-R (20) reliabilities

however, show a small but consistent advantage for Rights over Formula. The SAT-verbal study also indicated that when two tests were both given with Rights directions and scoring or when both were given with Formula directions and scoring, the correlations are slightly higher than when Formula directions and scoring are used for one test and Rights directions and scoring are used for the other.

On the important question of parallelism, the data from the SAT-verbal study yield estimated true-score coefficients that are close enough to unity to dispel any concerns that changing directions results in a change in the abilities measured. For GMAT, however, there is an indication that the abilities measured by quantitative tests may be affected slightly by changes in directions. It was judged, however, that these deviations from parallelism were not large enough to cause concern in score equating.

Finally, I would like to add that an important part of the study was concerned with score equating. Although Dr. Angoff and I decided not to describe this phase of the study in this symposium, our conclusion may be of interest. On the whole, it appears that equating based on the Invariance Hypothesis yields results that are in good agreement with those obtained by standard equating methods.

A Study of Hypotheses Basic to the Use of Rights and Formula Scores

William H. Angoff
and
William B. Schrader

Educational Testing Service
Princeton, New Jersey

AERA Meeting
New York, N.Y.

March 20, 1982

2

Guessing Directions for SAT-verbal
and for the Chemistry Achievement Test

The Rights directions for SAT-verbal were as follows:

> Read the directions below carefully, and answer the questions in this section according to these directions.
>
> Your score on this section will be based on the number of questions you answer correctly. No deduction will be made for wrong answers. You are advised to use your time effectively and to mark the best answer you can to every question, regardless of how sure you are of the answer you mark.

The Formula directions were essentially the directions used for operational administrations of SAT-verbal, as follows:

> Read the directions below carefully, and answer the questions in this section according to these directions.
>
> Students often ask whether they should guess when they are uncertain about the answer to a question. Your score on this section will be based on the number of questions you answer correctly minus a fraction of the number you answer incorrectly. Therefore, it is improbable that random or haphazard guessing will change your scores significantly. If you have some knowledge of a question, you may be able to eliminate one or more of the answer choices as wrong. It is generally to your advantage to answer such questions even though you must guess which of the remaining choices is correct. Remember, however, not to spend too much time on any one question.
>
> Do not worry if you are unable to finish this section or if there are some questions you cannot answer; many students leave questions unanswered. You should work as rapidly as you can without sacrificing accuracy. Do not waste time puzzling over a question which seems too difficult for you.

The special directions for the Chemistry examinees were precisely the same except that the word "test" was substituted for the word "section."

Table 1

Directions for Administering SAT-verbal

Form A

Group	Part 1	Part 2
1	Rights	Rights
2	Rights	Formula
3	Formula	Rights
4	Formula	Formula

Form B

Group	Part 1	Part 2
5	Rights	Rights
6	Formula	Formula

Directions for Administering the Chemistry Test

Group	Total Test
7	Rights
8	Formula

Table 2

Summary Statistics on Rights and Formula Scores[a]

SAT-verbal, Form A

Part	Group	Directions	No. of Cases	Rights Score Mean	Rights Score S.D.	Formula Score Mean	Formula Score S.D.
1	1&2	Rights	2094	22.18	6.82	17.06	8.19
1	3&4	Formula	2092	21.65	6.69	17.04	7.99
2	1&3	Rights	2080	18.24	6.37	12.26	7.70
2	2&4	Formula	2106	17.46	6.03	12.85	7.32
Total	1	Rights	1026	40.47	12.75	30.30	15.36
Total	4	Formula	1038	38.99	11.90	29.70	14.36

SAT-verbal, Form B

Part	Group	Directions	No. of Cases	Rights Score Mean	Rights Score S.D.	Formula Score Mean	Formula Score S.D.
Total	5	Rights	1040	40.30	13.07	30.08	15.82
Total	6	Formula	1034	39.03	12.86	29.76	15.66

Chemistry Test

Part	Group	Directions	No. of Cases	Rights Score Mean	Rights Score S.D.	Formula Score Mean	Formula Score S.D.
Total	7	Rights	1151	34.59	10.74	22.06	13.03
Total	8	Formula	1155	33.21	10.26	21.52	11.90

[a]Statistics for variables for which the scoring is consistent with the directions are underlined.

4

Table 3

Effect of Differences in Directions on Formula Scores Earned on Part 2 of SAT-verbal for Groups Stratified on the Basis of Rights Scores, Formula Scores, and Nonattempts on Part 1[a]

Stratification Variable	Interval of Score on Part 1	Group Tested Using Rights Directions	Group Tested Using Formula Directions	t-test Results		
				N	Adjusted Difference on Part 2[b]	t[c]
Rights Score	25 and higher	1	2	734	+0.75	+1.69
"	20-24	"	"	600	+0.00	+0.00
"	15-19	"	"	412	+0.42	+0.91
"	14 and lower	"	"	262	+0.34	+0.63
"	Total	"	"	2008	+0.38	+1.41
Formula score	21 and higher	3	4	665	+0.27	+0.63
"	15-20	"	"	613	+0.29	+0.82
"	9-14	"	"	447	+0.37	+0.89
"	8 and lower	"	"	280	+0.72	+1.55
"	Total	"	"	2005	+0.60	+2.36*
Items Not Attempted	9 and higher	3	4	475	+0.67	+1.32
"	8 and lower	"	"	1530	+0.59	+2.00*
"	Total	"	"	2005	+0.60	+2.36*

[a]Covariates used in all analyses were: School, Sex, Rank in Class, and Ethnic Group. Students with missing data on any covariate were excluded.

[b]In all analyses, a plus sign indicates that students tested using Rights directions earned higher mean adjusted scores.

[c]Significance levels, using two-tailed tests:
* $p < .05$

Table 4

Effect of Differences in Directions on Formula Scores: Black Students[a]

Dependent Variable				Covariates	Group(s) Tested Using Rights Directions	Group(s) Tested Using Formula Directions	t-test Results		
Test	Form	Part(s)	Items				N	Adjusted Difference[b]	t
SAT-V	A	1	45	School, Sex, Rank	1+2	3+4	760	-0.23	-0.52
"	A	2	40	"	1+3	2+4	760	-0.10	-0.27

[a]Students with missing data on any covariate were excluded.

[b]In all analyses, a plus sign indicates that students tested using Rights directions earned higher mean adjusted scores. A minus sign indicates that students tested using Formula directions earned higher mean adjusted scores.

Table 5

Effect of Directions on Part 1 Scores on Formula Scores Earned on Part 2

Covariates	Groups Studied	No. of Cases	Adjusted Difference[a]	t[a]
Based on all students in designated groups				
Sch, Sex, Read, Ed.Gp.	1 and 3	1992	-0.13	-0.50
"	2 and 4	2021	-0.11	-0.42
Based on Black students in designated groups				
Sch, Sex	1 and 3	389	-0.97	-1.82
"	2 and 4	371	-0.23	-0.42

[a] In this table a minus sign indicates that students who had the same directions on both parts of the test earned lower mean adjusted scores on Part 2 than did those students who had different directions on the two parts of the test.

Table 6

Number of Items Omitted, Not Reached, Not Attempted, and Guessed on Part 2 of SAT-V, Given with Rights Directions, for White (W), Black (B) and Black plus Hispanic (B+H) Students, Stratified by Rights Scores on Part 1

(Based on Group 1)

Score Interval	Ethnic Group	N	Omitted Mean	Omitted SD	Not Reached Mean	Not Reached SD	Not Attempted[a] Mean	Not Attempted[a] SD	Guessing Index (W-O) Mean	Guessing Index (W-O) SD
25 and higher	W	316	0.25	1.20	0.34	1.29	0.59	1.78	15.03	5.35
	B	18	0.22	0.55	0.33	1.03	0.56	1.10	17.89	4.59
	B+H	23	0.26	0.54	0.26	0.92	0.52	0.99	18.22	5.13
20 - 24	W	223	0.66	2.10	0.58	1.62	1.25	2.71	20.20	5.68
	B	46	1.07	2.67	0.80	1.98	1.87	3.32	20.78	6.33
	B+H	52	0.96	2.53	0.83	2.02	1.79	3.25	20.69	6.09
15 - 19	W	120	0.80	2.75	1.05	2.19	1.85	3.55	22.22	6.63
	B	61	0.48	2.45	1.61	3.21	2.08	4.59	24.38	6.74
	B+H	79	0.38	2.16	1.34	2.92	1.72	4.13	24.75	6.23
14 and lower	W	35	2.40	5.80	1.37	2.74	3.77	6.44	21.49	11.67
	B	77	1.39	3.46	3.01	4.77	4.40	7.12	24.53	9.41
	B+H	90	1.39	3.39	2.87	4.59	4.26	6.78	24.29	9.16
Total	W	694	0.59	2.29	0.59	1.70	1.18	2.93	18.26	6.84
	B	202	0.94	2.85	1.85	3.69	2.78	5.46	23.04	7.92
	B+H	244	0.86	2.70	1.69	3.51	2.56	5.14	23.10	7.63

[a] Omits plus Not Reached

6

Table 7

Number of Items Omitted, Not Reached, Not Attempted, and Guessed on Part 2 of SAT-V, Given with Formula Directions, for White (W), Black (B), and Black plus Hispanic (B+H) Students, Stratified by Rights Scores on Part 1

(Based on Group 2)

Score Interval	Ethnic Group	N	Omitted Mean	Omitted SD	Not Reached Mean	Not Reached SD	Not Attempted[a] Mean	Not Attempted[a] SD	Guessing Index (W-O) Mean	Guessing Index (W-O) SD
25 and higher	W	341	3.32	3.94	0.43	1.34	3.75	4.29	10.26	8.01
	B	26	3.50	4.18	0.69	2.11	4.19	4.72	12.65	8.99
	B+H	30	3.43	4.17	0.77	2.13	4.20	4.96	12.43	8.48
20 - 24	W	233	3.47	4.38	0.94	1.99	4.41	5.19	14.79	8.98
	B	43	2.44	4.50	2.05	2.86	4.49	6.03	16.58	9.10
	B+H	56	3.21	5.20	1.80	2.69	5.02	6.27	15.64	10.30
15 - 19	W	120	3.19	4.71	1.55	2.84	4.74	5.73	17.93	9.97
	B	57	1.86	3.25	1.65	3.22	3.51	5.35	22.25	7.95
	B+H	70	2.03	3.53	1.67	3.14	3.70	5.35	21.56	8.02
14 and lower	W	49	3.45	5.20	2.53	3.38	5.98	6.25	19.45	10.53
	B	65	1.52	3.60	2.40	3.59	3.92	5.44	24.02	8.71
	B+H	78	1.50	3.71	2.18	3.39	3.68	5.38	24.27	8.60
Total	W	743	3.36	4.29	0.91	2.11	4.27	5.00	13.53	9.41
	B	191	2.10	3.83	1.86	3.18	3.96	5.43	20.27	9.50
	B+H	234	2.32	4.18	1.76	3.03	4.07	5.54	19.88	9.81

[a]Omits plus Not Reached

Table 8

Effect of Directions on Guessing Behavior

Test	Form	Group	Directions	No. of Cases	Mean 40 (Parts 1+2)
SAT-verbal	A	1	Rights	1026	39.28
"	"	4	Formula	1038	31.61
"	B	5	Rights	1040	39.56
"	"	6	Formula	1034	31.35
Chemistry	—	7	Rights	1151	47.34
"	—	8	Formula	1155	32.04

Table 9

Correlations between Responses
on the Two Parts of SAT-verbal, Form A

Group	Directions on the Two Parts	No. of Cases	Correlations between Parts 1 and 2		
			Rights	Wrongs	Omits
1	Same (R;R)	1026	.822	.768	.784
2	Different (R;F)	1068	.788	.686	.246
3	Different (F;R)	1054	.806	.695	.318
4	Same (F;F)	1038	.819	.807	.770

Table 10

Observed-Score Correlations between Pt-Scores[a]

Group	No. of Cases	Directions Part 1	Directions Part 2	$r_{R_1R_2}$	$r_{R_1F_2}$	$r_{F_1R_2}$	$r_{F_1F_2}$
			Form A				
1	1026	Rights	Rights	.822	.813	.816	.818
2	1068	Rights	Formula	.788	.798	.785	.804
3	1054	Formula	Rights	.806	.801	.810	.815
4	1038	Formula	Formula	.819	.809	.811	.824
			Form B				
5	1040	Rights	Rights	.820	.809	.823	.820
6	1034	Formula	Formula	.826	.815	.829	.837

[a] Correlations between variables for which the scoring is consistent with the directions are underlined.

Table 11

Reliability Coefficients[a]

Group	No. of Cases	Directions: Part 1	Directions: Part 2	r_{t,R_1}[b]	r_{t,A_1}[c]	r_{t,R_2}[b]	r_{t,R_2}[c]
			Form A				
1	1024	Rights	Rights	.848	.838	.838	.821
2	1068	Rights	Formula	.838	.830	.809	.809
3	1054	Formula	Rights	.845	.840	.822	.812
4	1038	Formula	Formula	.831	.825	.803	.802
			Form B				
5	1040	Rights	Rights	.849	.837	.844	.841
6	1034	Formula	Formula	.846	.842	.842	.844

[a] Correlations between variables for which the scoring is consistent with the directions are underlined.

[b] Kuder-Richardson formula (20) reliability.

[c] Dressel (1944) adaptation of KR (20) reliability.

Table 12

True-Score Correlations

Group	No. of Cases	Directions: Part 1	Directions: Part 2	r_{t,R_1}	r_{t,R_2}	r_{t,R_1}	r_{t,R_2}
			Form A				
1	1026	Rights	Rights	.991	.975	.989	.986
2	1068	Rights	Formula	.954	.970	.958	.982
3	1054	Formula	Rights	.967	.967	.975	.987
4	1038	Formula	Formula	1.003	.991	.997	1.013
			Form B				
5	1040	Rights	Rights	.968	.957	.979	.977
6	1034	Formula	Formula	.978	.965	.984	.992

[a] Correlations between variables for which the scoring is consistent with the directions are underlined.

10

TEST DIRECTIONS: GMAT STUDY

The following statement was printed at the beginning of tests given with Rights directions:

> Before answering the questions in this section, please read carefully the following directions, which apply only to this section.
>
> Your score on this test will be based on the number of questions you answer correctly. No deductions will be made for wrong answers. You are advised to use your time effectively and to mark the best answer you can to every question, regardless of how sure you are of the answer you mark.

The following statement was printed at the beginning of tests given with Formula directions:

> Students often ask whether they should guess when they are uncertain about the answer to a question. Your score on this section will be based on the number of questions you answer correctly minus a fraction of the number you answer incorrectly. Therefore, it is improbable that random or haphazard guessing will change your score significantly. If you have some knowledge of a question, you may be able to eliminate one or more of the answer choices as wrong. It is generally to your advantage to answer such questions even though you must guess which of the remaining choices is correct. Remember, however, not to spend too much time on any one question.

Table 13

TEST DESIGN: GMAT STUDY

OPERATIONAL TEST

(These six sections were identical on all ten forms)

Section	Item Type	Directions	Number of Items	Time in Minutes
1	Reading Comprehension	F	25	30
2	Problem Solving	F	30	40
3	Practical Business Judgment	F	20	20
4	Data Sufficiency	F	30	30
5	Practical Business Judgment	F	20	20
7	Usage	F	25	15

EXPERIMENTAL TESTS

(These tests were given as Section 8 on all ten forms)

Form	Item Type	Directions	Number of Items	Time in Minutes
1	Reading Comprehension	F	29	30
2	Reading Comprehension	R	29	30
3	Problem Solving	F	25	30
4	Problem Solving	F	25	30
5	Practical Business Judgment	F	32	30
6	Practical Business Judgment	R	32	30
7	Data Sufficiency	F	40	30
8	Data Sufficiency	R	40	30
9	Sentence Correction	F	30	30
10	Sentence Correction	R	30	30

12

Table 14

Descriptive Statistics on Rights and Exact (Unrounded) Formula Scores for GMAT Experimental Tests

Experimental Test	Number of Items	Directions	N	Rights Score[a] Mean	Rights Score[a] S.D.	Formula Score[a] Mean	Formula Score[a] S.D.
Reading Comprehension	29	R	5658	*15.03*	*6.21*	12.35	7.11
Reading Comprehension	29	F	5739	14.68	6.46	*12.38*	*7.17*
Problem Solving	25	R	5501	*9.87*	*4.04*	7.56	4.52
Problem Solving	25	F	5594	9.00	4.02	*7.57*	*4.43*
Practical Business Judgment	32	R	5738	*18.82*	*4.84*	15.67	5.87
Practical Business Judgment	32	F	5408	18.95	4.87	*15.85*	*5.89*
Data Sufficiency	40	R	5590	*18.38*	*5.33*	12.31	5.91
Data Sufficiency	40	F	5657	18.18	5.46	*12.42*	*5.92*
Sentence Correction	30	R	5409	*17.05*	*5.24*	14.44	6.07
Sentence Correction	30	F	5486	16.79	5.40	*14.43*	*6.03*

[a]Means and standard deviations for which the scoring is consistent with the directions appear in italics.

Table 15

Descriptive Statistics on Number of Items Omitted, Not Reached, and Not Attempted for GMAT Experimental Tests

Experimental Test	Number of Items	Directions	N	Omitted Mean	Omitted S.D.	Not Reached Mean	Not Reached S.D.	Not Attempted Mean	Not Attempted S.D.
Reading Comprehension	29	R	5658	1.32	2.67	1.94	3.94	3.26	4.89
Reading Comprehension	29	F	5739	2.44	3.30	2.66	4.35	5.09	5.35
Problem Solving	25	R	5501	4.76	4.84	2.15	3.51	6.90	5.71
Problem Solving	25	F	5594	7.41	4.36	2.90	3.84	10.31	4.41
Practical Business Judgment	32	R	5738	0.44	1.65	0.13	1.10	0.57	2.14
Practical Business Judgment	32	F	5408	0.50	1.72	0.14	1.10	0.64	2.22
Data Sufficiency	40	R	5590	4.31	5.04	3.02	4.78	7.33	6.66
Data Sufficiency	40	F	5657	5.62	5.43	3.18	4.89	8.80	6.69
Sentence Correction	30	R	5409	1.03	2.43	1.47	3.06	2.50	3.92
Sentence Correction	30	F	5486	1.73	2.98	2.05	3.54	3.78	4.49

Table 16

Number of Items Omitted, Not Reached, Not Attempted, and Guessed on the Experimental Section of Reading Comprehension, Given with Rights Directions, Stratified by Formula Scores on the Operational Section of Reading Comprehension

Operational Test Score Interval	No. of Cases	Omitted		Not Reached		Not Attempted		Guessing Index (W-O)	
		Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
22-25	189	0.29	0.82	0.25	0.89	0.54	1.42	5.14	3.05
20-21	529	0.46	1.33	0.56	1.87	1.02	2.43	6.12	3.54
18-19	835	0.59	1.54	0.71	2.05	1.30	2.71	7.69	4.12
16-17	580	0.93	1.93	0.98	2.36	1.91	3.19	8.34	4.67
14-15	879	1.09	2.16	1.48	3.18	2.57	3.96	9.49	5.15
12-13	640	1.61	2.73	2.10	3.90	3.71	4.72	9.75	6.03
10-11	650	1.78	3.00	2.40	4.17	4.19	5.01	10.92	6.33
8- 9	514	1.73	2.87	3.15	4.91	4.90	5.50	11.84	6.41
6- 7	339	2.71	4.13	3.38	4.84	6.09	6.09	10.67	8.17
4- 5	257	2.07	3.44	4.22	5.62	6.28	6.42	12.69	7.68
2- 3	124	2.21	3.48	6.79	7.02	9.00	7.80	12.02	9.05
0- 1	122	2.95	5.08	4.69	6.04	7.64	7.49	12.66	10.19
Total	5658	1.32	2.67	1.94	3.94	3.26	4.89	9.39	6.12

Table 17

Number of Items Omitted, Not Reached, Not Attempted, and Guessed on the Experimental Section of Reading Comprehension, Given with Formula Directions, Stratified by Formula Scores on the Operational Section of Reading Comprehension

Operational Test Score Interval	No. of Cases	Omitted		Not Reached		Not Attempted		Guessing Index (W-O)	
		Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
22-25	186	0.70	1.44	0.72	2.20	1.42	2.92	4.05	3.02
20-21	586	0.95	1.82	0.65	1.90	1.60	2.72	5.18	3.46
18-19	879	1.23	2.01	0.99	2.44	2.22	3.25	6.11	4.06
16-17	568	1.90	2.50	1.86	3.30	3.77	4.15	6.04	4.72
14-15	845	2.10	2.75	2.16	3.62	4.26	4.53	7.20	5.21
12-13	638	2.93	3.04	3.08	4.27	6.01	4.61	6.92	5.28
10-11	681	3.16	3.59	3.21	4.46	6.37	5.15	7.75	6.34
8- 9	519	3.44	3.83	4.21	5.00	7.65	5.55	7.90	7.17
6- 7	325	4.28	4.41	5.30	5.52	9.58	5.61	6.70	7.30
4- 5	253	4.06	4.29	5.16	5.75	9.22	6.23	8.44	7.85
2- 3	142	4.63	4.53	6.20	6.23	10.83	6.16	7.68	7.68
0- 1	117	4.03	5.21	6.34	7.26	10.37	7.45	9.15	9.38
Total	5739	2.44	3.30	2.66	4.35	5.09	5.35	6.79	5.73

14

Table 18

Effect of Directions on Guessing Behavior

Experimental Test	Experimental Test Directions	Number of Cases	Mean Guessing Index (W-O)
Reading Comprehension	R	5658	9.39
	F	5739	6.79
Problem Solving	R	5501	3.67
	F	5594	1.71
Practical Business Judgment	R	5738	12.17
	F	5408	11.91
Data Sufficiency	R	5590	11.97
	F	5657	9.39
Sentence Correction	R	5409	9.42
	F	5486	7.70

Table 19

Reliabilities of Operational and Experimental Tests, and Raw Score and True-Score Correlations of Operational and Experimental Tests[a]

Test	Directions for Exp. Tests	Reliability Coefficients: Oper. Scored Formula	Reliability Coefficients: Exp. Scored Rights	Reliability Coefficients: Exp. Scored Formula	Correlation of Exp. with Oper. Raw Scores: Exp. Scored Rights	Correlation of Exp. with Oper. Raw Scores: Exp. Scored Formula	Correlation of Exp. with Oper. True Scores: Exp. Scored Rights	Correlation of Exp. with Oper. True Scores: Exp. Scored Formula
Reading Comprehension	R	.766	.854	.836	.728	.726	.900	.907
	F	.771	.868	.848	.733	.732	.896	.905
Problem Solving	R	.742	.719	.686	.687	.709	.940	.994
	F	.752	.736	.720	.735	.740	.988	1.006
Practical Business Judgment-Section 3	R	.625	.748	.736	.593	.589	.868	.869
	F	.631	.752	.737	.595	.592	.864	.868
Practical Business Judgment-Section 5	R	.588	.748	.736	.543	.536	.818	.816
	F	.586	.752	.737	.559	.553	.842	.840
Data Sufficiency	R	.812	.751	.710	.676	.716	.866	.944
	F	.819	.768	.722	.695	.731	.877	.950
Usage/Sentence Correction	R	.755	.782	.758	.670	.677	.873	.895
	F	.750	.796	.763	.672	.677	.869	.894

[a]Coefficients based on scoring that is consistent with the directions are underlined.